IJDC | *Conference Paper*

Resolving Conflicts in Data Through Curation-informed Weight Distribution Networks

Nevan Simone
The University of Texas at Austin

Maria Esteva
The University of Texas at Austin

Abstract

Missing and conflicting data values create problems when integrating datasets from multiple collections. Moreover, when the collections to be integrated are large and continuously updated, it is not feasible to manually resolve these problems. Instead, disagreements and gaps should be resolved in an automated fashion. To achieve good quality integrated datasets automatically we introduce the Curation-informed Weight Distribution Network (CiWDN), a method that suggests which collection is more reliable in providing a data value in question. CiWDN adapts the PageRank algorithm (PR) to assign and distribute weights across data fields present in the different collections. Weight assignment is rooted in data curation best practices as metrics of a collection's reliability. The metrics include: a) data completeness, b) data coincidence, and c) data consistency over time. Final weights used as collection ranks provide the basis to resolve conflicts between different collections contributing a data value for a given field. CiWDN relies on a data dictionary that normalizes fields across collections, and is implemented on a graph database. We demonstrate CiWDN's capability using the case of ASTRIAGraph, a knowledge system built to increase transparency of activities in Earth's orbital environment. CiWDN can assess the reliability of data collections that conflict on space object characteristic data fields, which can be used to resolve the differences. This method for computing collections' reliability can be ported to curate other types of large integrated datasets for use in machine learning and other data-driven applications.

Submitted 9 February 2024 ~ *Accepted* 20 February 2024

Correspondence should be addressed to Nevan Simone, 201 E 24th St, Austin, TX 78712. Email: nfs342@utexas.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v18i1.935
DOI: 10.2218/ijdc.v18i1.935

Introduction

Data curation entails activities towards maintaining, preserving, and assuring the quality of research data throughout the continuum of its generation and reuse (Pouchard, 2015). It is often characterized as labor-intensive and skill specific (Borgman & Bourne, 2022), frequently requiring ad-hoc solutions for each dataset (Thirumuruganathan et al., 2020). Such is the case for domain specific datasets that integrate data from multiple collections, each of which may provide different, overlapping, or missing values for the same data field. Problems are compounded when contributing collections are very large, grow over time, and data values are constantly updated, in which case detecting and resolving conflicts manually becomes untenable. We introduce Curation-informed Weight Distribution Network (CiWDN), a novel method to automatically provide a recommendation as to which collection provides the best value for a data field with conflicting values.

An important concept in this work is reliability, which we define as the confidence that can be placed on the data provided by a given collection. Reliability is akin to "trustworthy," "with integrity," or "reproducible," all terms used to express data curation goals. Translated as different aspects of a data collection's quality, reliability is assessed using established data curation best practices (Sawchuk, Gillis, & Macleod, 2023). In this work collections' reliability is automatically assessed through three metrics: a) completeness, b) coincidence, and c) consistency. These metrics are applied as weights in the form of punishment or rewards to come up with a collection's reliability score.

Implemented in a graph database, CiWDN uses the PageRank algorithm (Page, 2006) as a mechanism to assess data fields and corresponding values within and across collections, distribute weights according to the established metrics, and rank the reliability of the collections being compared. The methodology relies on previous work done by this team designing a unifying data model and corresponding data dictionary so that fields across collections are interoperable (Esteva et al., 2022). Based on the model, field data labels from different collections are normalized during ingestion to the graph database.

We test CiWDN using ASTRIAGraph1, a knowledge system that is used to answer questions about space sustainability. This system has complex curation challenges in that it has aggregated data about satellites and other man-made space objects (here referred to as SOs) from different collections since 2018. These collections contribute a variety of fields, and each is updated at a different pace over time. CiWDN helps determine the most reliable collection to provide the value of each field, which offers a solution for SOs resolution when there is no agreement between collections.

Related Work

Applications required to conduct data curation tasks automatically or semi-automatically are increasingly being developed (Thirumuruganathan et al., 2020). In their survey of 667 data quality tools, Ehrlinger and Wöß (2022) found the pool to be diverse in terms of scope and functions, but most tools lacked rule-based metrics and data quality metrics reporting. In their framework for digital curation Yoon et al. propose that curation assurance, understood as trust in the data, is an area in need of development (Yoon et al., 2022). The method proposed here fills the gaps identified by these two papers. Curation is automated by operationalizing best practices against which the reliability of data collections is measured.

Ehrlinger and Wöß (2022) also found that 50% of the tools surveyed are domain specific. Many such tools require semantic models and data mappings to adjust to the requirements of

[1] ASTRIAGraph: http://astria.tacc.utexas.edu/AstriaGraph/

domain specific datasets (Malik, et al., 2020; Satti, et al., 2020). Our methodology is both domain specific and generalizable. As long as a data dictionary is in place to address specific data integration needs, the curation metrics and weight distribution methodology can be ported to data in any domain.

A recurrent theme in domain specific data curation is the required level of human effort. IQBot was designed to share the specialized work done by curators in protein databases that cannot afford them (Alqasab et al., 2017). The algorithm detects errors between current and previous versions of data in databases that use extensive human curation expertise. The detection results are then ported to other protein databases for purposes of comparison and error detection. The authors suggest that the code can be applied to other than protein datasets. While our method compares with other collections to suggest the best values to resolve a given conflict, suggestions are not based on the expertise of a curator but on the degree of reliability of the collections that is computed automatically by CiWDN.

Entity resolution and collection reliability are the core of this work. Data Tamer uses artificial intelligence to merge collections with different formats and field names (Stonebraker et al., 2013). With the goal of reducing the number of resolution cases by orders of magnitude, it detects when entries reference the same object, and it leaves cases with too much ambiguity for resolution by human input. SLiMFast, which focuses on research articles, approaches data resolution as a statistical learning problem (Rekatsinas et al., 2017). The method quantifies the reliability of each article by comparing its results to the rest of the data in the database, considering the likelihood that the article is correct based on other entries. The system also allows users to label particular entries as “truth.” The tool described by Fourches, Muratov, and Tropsha (2016) curates chemical data and bioprofiles to flag errors in chemogenomics research. Focused on data quality, the tool provides a measurement for research results that can be verified or refuted through repeated experiments. Our project differs from these systems in that entity resolution is based on curation metrics and is implemented through an algorithm that ranks collections reliability based on their contents, without human feedback or experiment validation.

The CiWDN Method

CiWDN is built on an adaptation of PageRank (PR). PR is an algorithm designed to rank related entries in linked databases. Each entry is assigned a weight, interpreted as importance, and the algorithm distributes weight iteratively from an entry to all others that it references until each entry’s weight numerically converges. The end result is a list of weights that ranks the entries in the database. Notably, Google used PR to rank the importance of websites related to a search query. Each site is an entry in Google’s World Wide Web database. A hyperlink from one page to another constitutes a link between those related sites. More accrued weight is interpreted as greater importance, which determines the order of search results.

Instead of entries, in CiWDN we adapted PR in a graph database to rank data collections where there is no guarantee that data field values agree between them. In this approach, the database is cast as a graph, where fields are nodes, and the relationships between nodes are directed links between them. The PR based method to assign weights to the nodes is informed by curation metrics of completeness, coincidence, and consistency.

To illustrate the application's capabilities, in Table 1 we show the characteristics and scope of the 7 collections that provide data to ASTRIAGraph. The names of the collections have been anonymized through an ID, each row showing the number of SOs and corresponding fields that they report on as well as the frequency with which the data is updated. Note that not all field values change at every update.

Table 1. Scope of Collections Providing Data About SOs to ASTRIAGraph

Collection ID	Number of SOs	Number of Fields	Collection Update Frequency
0	23983	20	daily
1	352	18	daily
3	7081	16	daily
4	210	18	daily
7	4542	15	semi-monthly
13	669	13	single import
15	733	17	single import

Data provided by heterogeneous collections can be categorized as static and dynamic. Static fields contain the same values throughout different data updates, while dynamic fields contain values that change during updates over time. Also, some fields may contain values with literal meaning, such as a name in a string format. Others may be domain-specific formats with information encoded through scientific algorithms. Dynamic fields, especially for ASTRIAGraph, vary numerically, and the meaning in the variations is dependent on the context of each field. Thus, assessment requires computing similarity between values using domain-specific algorithms. To demonstrate the generalizability of the method, this work focuses on static fields.

Adaptation of the PR algorithm in the ASTRIAGraph database is as follows. First, collections ingested to the database go through normalization of field labels. This allows the weight assignment process to go through the graph database nodes within and across collections (Esteva et al., 2020). A PR node is defined as a collection's field, linked to other related fields, and encompassing the field's current and historical data values. The weights of nodes are set according to curation metrics that will indicate the quality of that node's data. Figure 1 depicts how weights are distributed across nodes.

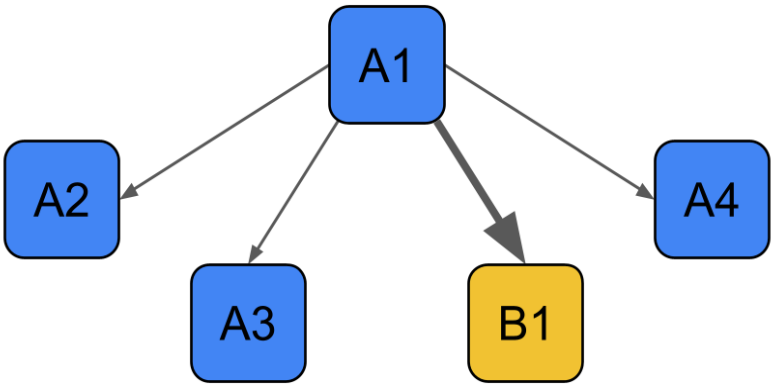

Figure 1. Arrows of different sizes indicate the scale at which weight is distributed from the source node to other nodes in the collection (blue) or to a different collection (yellow).

Each node will have links pointing to other nodes. Each pair of nodes in the same collection will have links between them. A link will point outside its collection if there is a similar node in another collection. The distributed weight is split across these links and given to the destination nodes, but the partitioning of weight to links is not necessarily equal between the links. Each link is assigned an attribute to determine how much weight is passed from its origin node during weight distribution. We call this attribute a “scaling factor." The three curation metrics – completeness, coincidence, and consistency – determine the initial weights of the nodes and the scaling factors of links across collections. These three metrics represent patterns of gaps, changes, conflicts, and commonalities in data values within and across collections.

Represented in Figure 2, completeness evaluates the presence of gaps within the most recent data in a collection’s node. Gaps can be blank values or strings denoting an unknown value such as “TBD.”

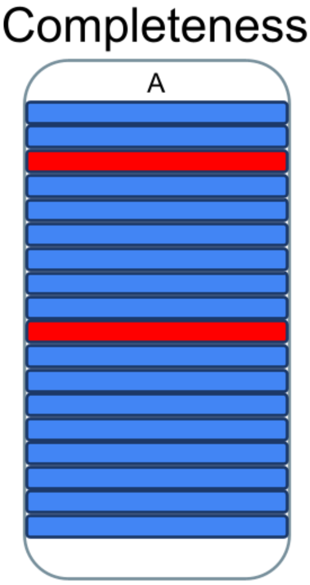

Figure 2. Red blocks represent gaps in collection A.

Coincidence compares data values between two similar nodes in different collections. We compute the percentage of values in one node’s present in the other node’s data (see Figure 3).

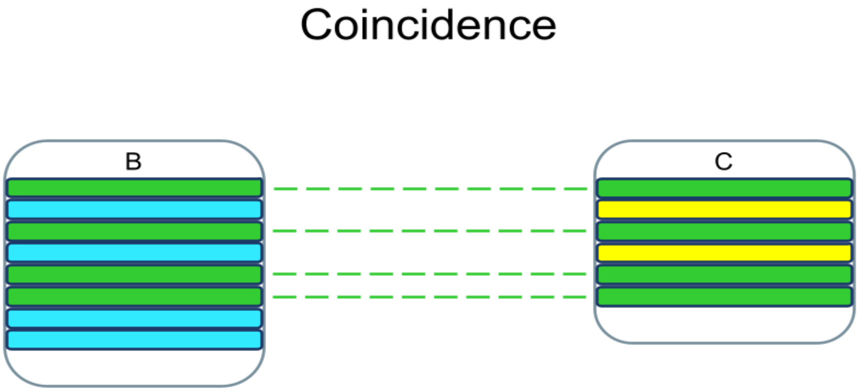

Figure 3. Green blocks represent common data values between nodes in collections B and C.

The coincidence result is used to scale the weight distributed between node B and node C. Larger coincidence increases the amount of weight to be distributed.

Consistency considers the historical data corresponding to each node. By historical data, we refer to data that aggregates over time through updates noted in Table 1. We chose to sample each node over the span of a year on intervals of every tenth day. Between concurrent samples, we look for changes, whether new gaps or altered values, as shown in Figure 4. As such, this metric combines completeness and coincidence within the historical data of a single node.

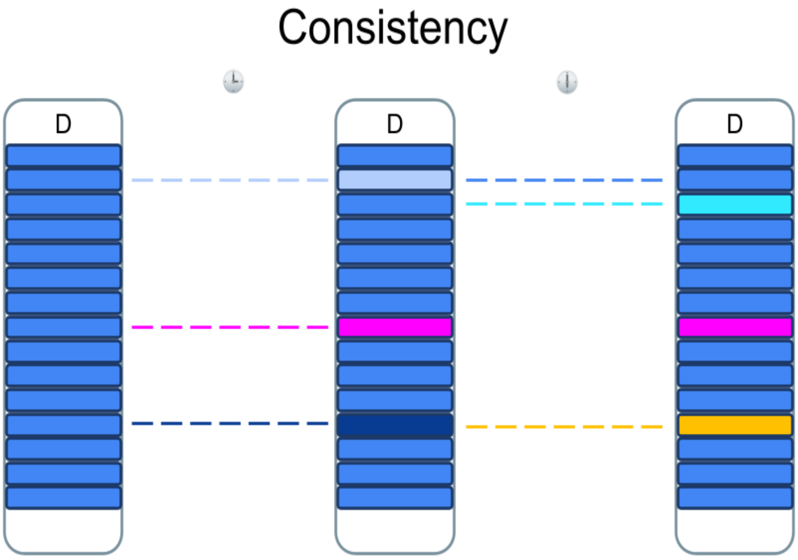

Figure 4. Different colors indicate values changing in a node in collection D over time.

Completeness and consistency metrics, which evaluate each node in isolation, are used to initialize all weights in the graph's nodes. After this, the PR-based weight assignment loop begins and each node is evaluated sequentially. Figure 5 shows how this works in three steps: 1) the node's weight is updated, 2) the weight is compared, and 3) the weight is distributed.

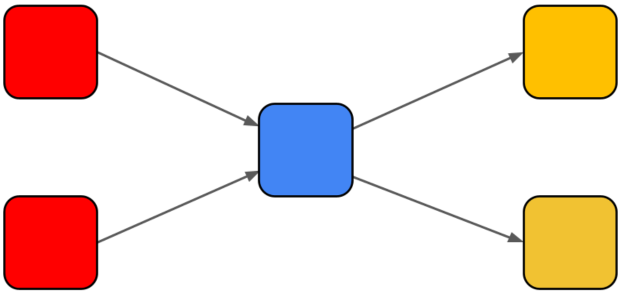

Figure 5. Red indicates the update step into blue. The previous and updated blue weights are compared. Weight is distributed to yellow.

To update a node's weight in step 1, all the weights on links pointing towards the node are summed. Step 2 compares the updated weight to the previous one on the same node. The distribution in step 3, divides the node's weight onto links that point away from the node. The portion of weight distributed to each link depends on the collection the link points to. If it is a different collection, then coincidence between the nodes determines the amount of weight. The process moves on to the next node in the sequence until there are no more iterations and the weights of all nodes converge. Final weights are the nodes' reliability scores, providing a basis for evaluating the quality of the collection field-by-field.

Demonstrating Collections Reliability

We illustrate how CiWDN delivers results through use cases involving the resolution of the static fields: Country (which refers to the country of origin of a SO) between collections 0, 1, and 7; and NORAD ID (an identifier system run by USSPACECOM and used worldwide) between collections 0, 1, 3, and 4.

First, we consider the resolution of the field Country. As shown in Table 2, while collection 0 contains a small percentage of gaps, and collections 0 and 7 contain a small number of changes across their historical data, the three can be considered complete and consistent.

Table 2. Country Completeness & Consistency of Collections 0, 1, 7.

Node Name	Completeness	Consistency
0_Country	0.989	0.999
1_Country	1.0	1.0
7_Country	1.0	0.991

The coincidence metric affects this case the most. Results shown in Table 3 indicate that collection 0 contains almost all of the Country values present in 1 and 7, while collection 1 contains almost no values from the other two. The coincidence scores distribute large portions of weight from 1_Country and 7_Country into 0_Country.

The metrics combined with the graph's node structure determine how weight is assigned and distributed. The results of the process are drawn in Figure 6, with the final reliability scores at the endpoints on the graph. This gives us a reliability ranking of these collections for field Country. Collection 0 is found to be the most reliable to resolve the Country field conflict. All three collections are rewarded roughly the same for their high completeness and consistency. The coincidence results reward collection 0 the most for containing many Country values from the other two collections. Collection 1 is the most punished by its low coincidence with the other collections, leaving Collection 7 with a moderate reliability score.

Table 3. Country Coincidence Between Collections 0, 1, and 7.

Compared Nodes/Fields	Coincidence
0_Country with 1_Country	9.901×10^{-3}
0_Country with 7_Country	9.901×10^{-3}
1_Country with 0_Country	1.0

Compared Nodes/Fields	Coincidence
1_Country with 7_Country	0.0
7_Country with 0_Country	1.389×10^{-2}
7_Country with 1_Country	0.0

The second case focuses on resolving NORAD ID between four collections. It illustrates how CiWDN can inform complex domain-specific decisions. Occasionally SOs orbit predictions from different collections will coincide, but the corresponding SOs identifiers may not match. This introduces ambiguity of whether different collections are reporting on the same or on two distinct SOs.

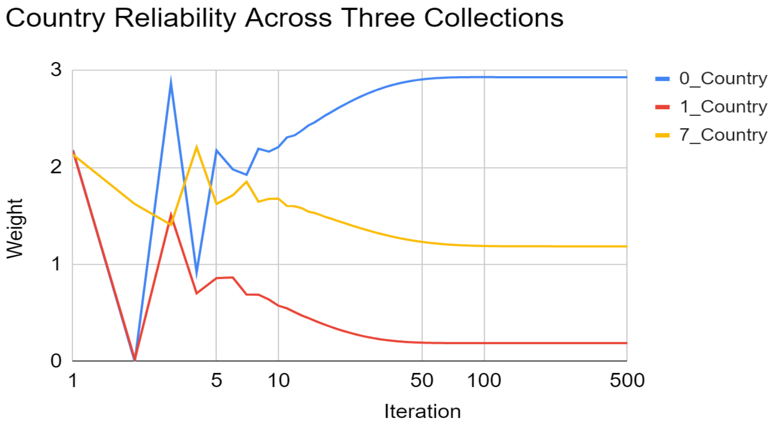

Figure 6. Country Reliability Scores for Collections 0, 1, and 7, demonstrating convergence over the weight assignment process.

Table 4 shows that the four collections are fairly complete and consistent for the NORAD ID field. Only collection 3 contains gaps, and the consistency scores indicate that few changes occur over time.

Table 4. NORAD ID Completeness & Consistency of Collections 0, 1, 3, 4.

Node Name	Completeness	Consistency
0_NoradId	1.0	0.996
1_NoradId	1.0	0.998
3_NoradId	0.999	0.943
4_NoradId	1.0	0.984

Again, coincidence affects the final reliability scores strongly. As shown in Table 5, coincidence in collection 0 is the highest across the board. Collections 1, 3, and 4 only contain a few common NORAD ID values. The coincidence scores distribute large portions of weight from these collections to collection 0, and very little weight from collection 0 to the others.

Table 5. NORAD ID Coincidence Between Collections 0, 1, 3, 4.

Compared Nodes/Fields	Coincidence
0_NoradId with 1_NoradId	1.468×10^{-2}
0_NoradId with 3_NoradId	2.264×10^{-2}
0_NoradId with 4_NoradId	8.339×10^{-5}
1_NoradId with 0_NoradId	1.0
1_NoradId with 3_NoradId	0.0
1_NoradId with 4_NoradId	0.0
3_NoradId with 0_NoradId	0.727
3_NoradId with 1_NoradId	0.0
3_NoradId with 4_NoradId	0.0
4_NoradId with 0_NoradId	9.524×10^{-3}
4_NoradId with 1_NoradId	0.0
4_NoradId with 3_NoradId	0.0

Using the reliability computed for the NORAD ID field, we would put more trust in the identifiers of collection 0, followed by collection 3, as shown in Table 6. In fact, the reliability for collection 0 is so high, that we would take its value, if available, even if there is agreement on a different value in other collections. The collections' reliability scores can help decide whether the identifier mismatch is due to an error, or to two SOs orbits nearing each other. These types of insights can be used to augment satellite tracking work.

Table 6. NORAD ID Final Reliability Scores.

Node Name	Reliability
0_NoradId	4.509
1_NoradId	0.258
3_NoradId	0.998
4_NoradId	0.581

We asked domain experts in the ASTRIAGraph team, who have purview on the sources of the data collections involved in the study, to qualitatively evaluate these results. They agreed with the results obtained, and indicated that collection 0 is a renowned, established collection, generated and used by a trustworthy and diligent organization with capable equipment.

Conclusion

We developed CiWDN to identify the reliability of collections providing data to a unified knowledge system. CiWDN informs choices between collections when field values do not match. Using ASTRIAGraph, we showed that the method can help identify the most reliable collection to resolve conflicting values. Based on an adaptation of the PR algorithm informed by three curation metrics, the method is particularly useful to resolve static data conflicts. Importantly, providing that a data model and dictionary are in place, CiWDN can be applied to resolve conflicts in data collections of diverse domains.

The main contribution of this methodology is that it addresses the reality of large-scale, multivariate, multi-collections data curation automatically, and establishes a valid path towards achieving transparency and improving the quality of data-driven science. The need for automated assessment of collection reliability is only increasing, as data aggregation and unification/merging of data collections are becoming commonplace in the face of AI. This work offers a solution based solely on the structure and contents of the data collections. While independent of human input, the results of the CiWDN can be considered by experts as a highly reliable recommendation for resolving conflicts.

To augment this system in future work, we will implement domain-specific algorithms to analyze dynamic values that are not trivially compared. Additionally, there are other edge cases. While this work considers each field within a collection individually, some fields come in sets, or can be partially derived from another field. We posit that CiWDN could be expanded to handle these cases. More complicated field relationships could be modelled with intermediate nodes added to the graph, accounting for the information embedded in the data. As identified by Ehrlinger and Wöß (2022), other metrics that measure data quality may be useful in knowledge system assessments. For example, using natural language processing to identify if words are misspelled or if two values are written differently but may have the same meaning. Since the methodology implements each curation metric independently, additional metrics could be added to expand CiWDN without disturbing this initial design. This makes CiWDN further generalizable to other knowledge systems.

References

Alqasab, M., Embury, S. M., & Sampaio, S. D. F. M. (2017). Amplifying data curation efforts to improve the quality of life science data. *International Journal of Digital Curation*, *12*, 1–12. Retrieved from ijdc.net. doi:10.2218/ijdc.v12i1.495

Borgman, C. L., & Bourne, P. E. (2021). Why it takes a village to manage and share data. *arXiv*. http://arxiv.org/abs/2109.01694

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, *5*. https://www.frontiersin.org/articles/10.3389/fdata.2022.850611

Esteva, M., Xu, W., Simone, N., Gupta, A., & Jah, M. (2020). Modeling data curation to scientific inquiry: A case study for multimodal data integration. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, *2020*. doi:10.1145/3383583.3398539

Esteva, M., Xu, W., Simone, N., Nagpal, K., Gupta, A., & Jah, M. (2022). Synchronic curation for assessing reuse and integration fitness of multiple data collections. *International Journal of Digital Curation*, *17(1)*. https://doi.org/10.2218/ijdc.v17i1.847

Fourches, D., Muratov, E., & Tropsha, A. (2016). Trust, but verify II: A practical guide to chemogenomics data curation. *Journal of Chemical Information and Modeling*, *56.7*, 1243-1252. doi:10.1021/acs.jcim.6b00129

Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications*, *145*, 113120. doi:10.1016/j.eswa.2019.113120

Page, L. (2006). *Method for Node Ranking in a Linked Database* (US 7058628 B1). US Patent and Trademark Office.

Pouchard, L. (2015). Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation*, *10.2*, 176–92. doi:10.2218/ijdc.v10i2.342

Rekatsinas, T., Joglekar, M., Garcia-Molina, H., Parameswaran, A., & Ré, C. (2017). SLiMFast: Guaranteed results for data fusion and source reliability. *Proceedings of the 2017 ACM International Conference on Management of Data*. doi:10.1145/3035918.3035951

Satti, F. A., Ali, T., Hussain, J., Khan, W. A., Khattak, A. M., & Lee, S. (2020). Ubiquitous Health Profile (UHPr): A big data curation platform for supporting health data interoperability. *Computing*, *102*, 2409–44. doi:10.1007/s00607-020-00837-2

Sawchuk, S., Gillis, L., & MacLeod, L. (2023). Supporting reproducible research with active data curation. *Research Data Management in the Canadian Context*. doi:10.5206/QQSG2445

Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A. & Xu, S. (2013). Data curation at scale: The Data Tamer system. *Conference on Innovative Data Systems Research, 2013*. https://cs.brown.edu/courses/csci2270/archives/2017/papers/data-tamer.pdf

Thirumuruganathan, S., Tang, N., Ouzzani, M., & Doan, A. (2020). Data curation with deep learning. *EDBT*, *1*. doi:10.5441/002/EDBT.2020.25

Yoon, A., Kim, J., & Donaldson, D. R. (2022). Big data curation framework: Curation actions and challenges. *Journal of Information Science*. doi:10.1177/01655515221133528